November 3, 2022

1 Task 1

1.1 Question 1.1

Q: What is the relationship between actor-critic and REINFORCE with baseline?

A: using value function as baseline with REINFORCE would basically make it an actor-critic method.

1.2 Question 1.2

Q: How can the value of advantage be intuitively interpreted?

A: Advantage can be interpreted as how much better the taken action is compared to the average action.

1.3 Question 1.3

Q: How does the implemented actor-critic method compare to REINFORCE in terms of bias and variance of the policy gradient estimation? Explain your answer.

A: Actor-critic model is more biased because the actions are valued based on the critic and not on monte-carlo estimates like with the policy-gradient method (REINFORCE). Thus it may take it more time to achieve same kind of level of performance than PG-methods, but once it does, it produces more stable results (less variance).

1.4 Question 1.4

Q: How could the bias-variance tradeoff in actor-critic be controlled?

A: Making the critic-network larger would decrease bias, but most likely increase variance.

1.5 Task 2

Training performance plot of deep deterministic policy gradient algorithm is depicted in figure 3.

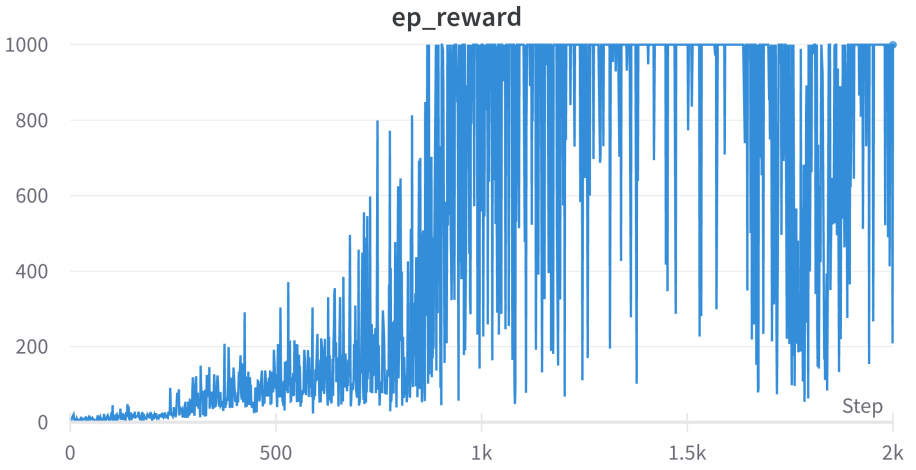

Figure 1: Training performance plot of actor-critic-method.

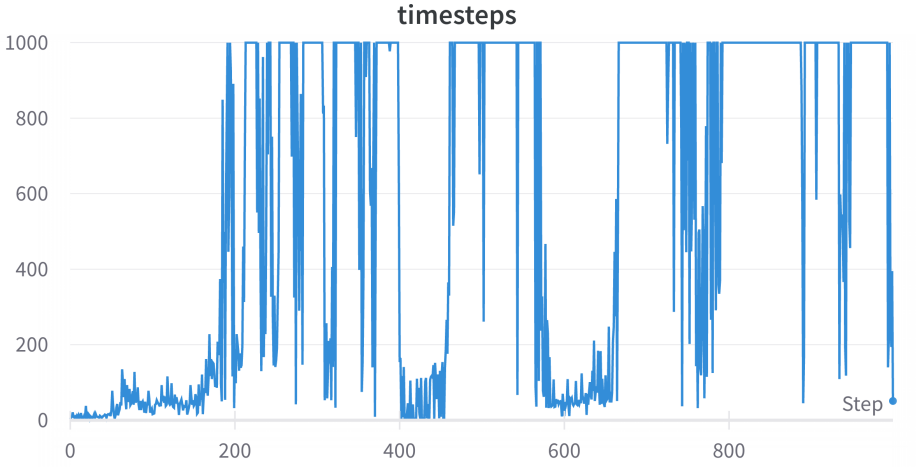

Figure 2: Training performance plot of REINFORCE with discounted rewards normalized to zero mean and unit variance and learnable parameter.

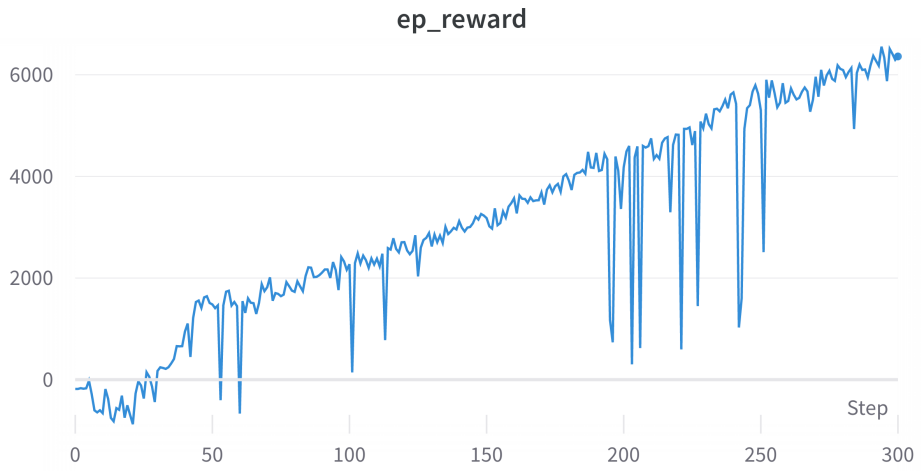

Figure 3: Training performance plot of deep deterministic policy gradient algorithm.

1.6 Question 2.1

Q: For policy gradient methods seen in Exercise 5, we update the agent using only on-policy data, while in DDPG we can use off-policy data. Why is this the case?

A: Because in DDPG we use replay buffer which makes it possible to use off-policy data. In policy gradient methods the use of replay buffer is not so easy.

1.7 Question 2.2

Q: A big advantage of DDPG is that it's able to utilise off-policy data. What are the disadvantages of deterministic policy gradient compared to the policy gradient method implemented in Task 1? List two of them.

A: Deterministic policy gradient method needs to learn critic, which runs the risk of bad hyperparametrization, where as in policy gradient the appraisal of policy is based on monte-carlo estimates of the reward, which in itself is unbiased.

Also stochastic policy gradient gives distributions for actions which can be sometimes beneficial over deterministic estimates.